

A Multilayer Approach to Multiplexity and Link Prediction in Online Geo-Social Networks

Desislava Hristova*, Anastasios Noulas*, Chloë Brown*, Mirco Musolesi, Cecilia Mascolo***

*Computer Laboratory, University of Cambridge

**School of Computer Science, University of Birmingham

dh475, an346, clb76, cm542 @ cam.ac.uk; m.musolesi @ cs.bham.ac.uk

Abstract

Online social systems are multiplex in nature as multiple links may exist between the same two users across different social networks. In this work, we introduce a framework for studying links and interactions between users beyond the individual social network. Exploring the cross-section of two popular online platforms - Twitter and location-based social network Foursquare - we represent the two together as a composite *multilayer online social network*. Through this paradigm we study the interactions of pairs of users differentiating between those with links on one or both networks. We find that users with multiplex links, who are connected on both networks, interact more and have greater neighbourhood overlap on both platforms, in comparison with pairs who are connected on just one of the social networks. In particular, the most frequented locations of users are considerably closer, and similarity is considerably greater among multiplex links. We present a number of structural and interaction features, such as the multilayer Adamic/Adar coefficient, which are based on the extension of the concept of the node neighbourhood beyond the single network. Our evaluation, which aims to shed light on the implications of multiplexity for the link generation process, shows that multilayer features, constructed from properties across social networks, perform better than their single network counterparts in predicting links across networks. We propose that combining information from multiple networks in a multilayer configuration can provide new insights into user interactions on online social networks, and can significantly improve link prediction overall with valuable applications to social bootstrapping and friend recommendations.

Introduction

Online social media has become an ecosystem of overlapping and complementary social networking services, inherently multiplex in nature, as multiple links may exist between the same pair of users (Kivelä et al. 2014). Multiplexity is a well studied property in the social sciences (Haythornthwaite and Wellman 1998) and it has been explored in social networks from Renaissance Florence (Padgett and Mclean 2006) to the Internet age (Haythornthwaite 2005). Despite the broad contextual differences, multi-relational ties are consistently found to exhibit greater intensity of interactions across different communication channels, and therefore a stronger

bond (Haythornthwaite and Wellman 1998; Hristova, Musolesi, and Mascolo 2014). Nevertheless, *there is a lack of research about online social networks and their value from a multiplex perspective*.

Recently, empirical models of multilayer networks have emerged to address the multi-relational nature of social networks (Kivelä et al. 2014; Szell, Lambiotte, and Thurner 2010). In such models, interactions are considered as layers in a systemic view of the social network. We adopt such a model in our analysis, where we shift the concept of a link and neighbourhood to encompass more than one network. This allows us to study interactions and structural properties across online social networks (OSNs), addressing the need for further understanding of their complimentary and overlapping nature, and multiplexity online. Although there have been some recent comparative studies of multiple online social networks (Ottino et al. 2014; Lee et al. 2014), and their intersection (Szell, Lambiotte, and Thurner 2010), the applications of multiplex network properties to OSNs is yet to be substantially addressed.

In this work, we explore intersecting networks, multiplex ties, and their application to link prediction across OSNs. Link prediction systems are key components of social networking services due to their practical applicability to friend recommendations and social network bootstrapping, as well as to understanding the link generation process. Link prediction is a well-studied problem, explored in the context of both OSNs and location-based social networks (LBSNs) (Liben-Nowell and Kleinberg 2007; Menon and Elkan 2011; Crandall et al. 2010; Scellato, Noulas, and Mascolo 2011). However, only very few link prediction works tackle multiple networks at a time (Lee et al. 2014; Tang, Lou, and Kleinberg 2012), while *most link prediction systems only employ features internal to the network under prediction*, without considering additional link information from other OSNs.

Our main contributions can be summarised as follows:

- We generalise the notion of a *multilayer online social network*, and extend definitions of neighbourhood to span multiple networks, adapting measures of overlap such as the Adamic/Adar coefficient in social networks to the multilayer context.

- We find that *pairs with links on both Twitter and Foursquare exhibit significantly higher interaction on both social networks* in terms of number of mentions and colocation within the same venues, as well as a lower distance and higher number of common hashtags in their tweets.
- A significantly higher overlap can be observed between the neighbourhoods of nodes with links on both networks, in particular with relation to the Adamic/Adar measure of neighbourhood overlap, which is significantly more expressed in the multilayer neighbourhood.
- In our evaluation, we predict Twitter links from Foursquare features and vice versa, and we achieve this with AUC scores up to 0.86 on the different datasets. In predicting links which span both networks, we achieve the highest AUC score of 0.88 from our multilayer features set, proving the multilayer construct a useful tool for social bootstrapping and friend recommendations.

The remainder of this work details these contributions, and summarises related work, concluding with a discussion of the implications, limitations, and applications of the proposed framework.

Related Work

Our work identifies with three main areas: multi-relational social networks, media multiplexity, and link prediction in online social networks. We summarise the state of the art in these areas in the following sections.

Multilayer Social Networks

Multi-relational or multilayer networks have been explored in the context of a wide range of systems from global air transportation (Cardillo et al. 2013) to massive online multiplayer games (Szell, Lambiotte, and Thurner 2010). A comprehensive review of multilayer network models can be found in (Kivelä et al. 2014). In the context of social networks, it is generally accepted that the more information we can obtain about the relationship between people, the more insight we can gain. A recent large-scale study on the subject has demonstrated the need for multi-channel data when comprehensively studying social networks (Stopczynski et al. 2014). Despite the observable multilayer nature of the composite OSNs of users (Kivelä et al. 2014; Kazienko et al. 2010; Bródka and Kazienko 2012), most research efforts have been focused on theoretical modelling (Kivelä et al. 2014), with little to no empirical work exploiting data-driven applications in the domain of multilayer OSNs, especially with respect to how location-based and social interactions are coupled in the online social space. We attempt to fill these gaps in the present work by presenting a generalisable online multilayer framework applied to classic problems such as link prediction in OSNs. Our framework is strongly motivated by the theory of media multiplexity, which we review next.

Media Multiplexity

Media multiplexity (Haythornthwaite 2005) is the principle that tie strength is observed to be greater when the num-

ber of media channels used to communicate between two people is greater (higher multiplexity). In (Haythornthwaite and Wellman 1998) the authors studied the effects of media use on relationships in an academic organisation and found that those pairs of participants who utilised more types of media (including email and videoconferencing) interacted more frequently and therefore had a closer relationship, such as friendship. More recently, multiplexity has been studied in light of multilayer communication networks, where the intersection of the layers was found to indicate a strong tie, while single-layer links were found to denote a weaker relationship (Hristova, Musolesi, and Mascolo 2014). The strength of social ties is an important consideration in friend recommendations and link prediction (Gilbert and Karahalios 2009), and we employ the previously understudied multiplex properties of OSNs to such ends in this work.

Link Prediction

The problem of link prediction was first introduced in the seminal work of Kleinberg et al. (Liben-Nowell and Kleinberg 2007) and since then, has been applied in various network domains. For instance, in (Scellato, Noulas, and Mascolo 2011) the authors exploit place features in location-based services to recommend friendships, and in (Backstrom and Leskovec 2011) a new model based on supervised random walks is proposed to predict new links in Facebook. Most of these works build on features that are endogenous to the system that hosts the social network of users. In our evaluation, however, we train and test on heterogeneous networks. In a similar spirit, the authors in (Sadilek, Kautz, and Bigham 2012) show how using both location and social information from the same network significantly improves link prediction. Our approach differs in that it frames the link prediction task in the context of multilayer networks and empirically shows the relationship between two different systems - Foursquare and Twitter - by mining features from both. Before presenting our framework and analysis, we will next state the research questions we are interested in answering through this work.

Research Questions

In light of the related work presented above, our goal is to mend the gap between multilayer network models, media multiplexity properties, and link prediction systems. More specifically, we address the following research questions in this work:

RQ1: *How do structural properties such as degree extend into the multilayer neighbourhood?* We propose a multilayer version of the network neighbourhood, which extends it to multiple networks (layers) and observe how such structural properties are manifested across Twitter and Foursquare.

RQ2: *What are the structural and behavioural differences between single network and multiplex links?* In order to understand the value of multiplex links (users connected on more than one network), we observe how they compare

to single network links in terms of neighbourhood overlap, Twitter interaction, similarity and mobility in Foursquare.

RQ3: *Can we use information about links from one layer to predict links on the other?* Many online social systems suffer from a lack of initial user adoption. Although many social networks nowadays incorporate the option of importing contacts from another pre-existing network and copying links, this method does not offer a ranking of users by relevance targeted towards the specific platform.

RQ4: *Can we predict links which exist on more than one network (i.e., multiplex links)?* Media multiplexity is a valuable source of tie strength information, and has further structural implications, which are of interest to OSN services and link prediction systems. We would like to explore the potential of identifying such links for building more successful online communities.

We will next present our multilayer framework for OSNs, and study user behaviour and properties across Twitter and Foursquare, extending our analysis to multiplex links in comparison with single-layer links. We finally integrate this into a link prediction system for OSNs, where we evaluate the utility of the metrics and features described in this work in hope to answer the above posed questions.

Multi-relational Framework

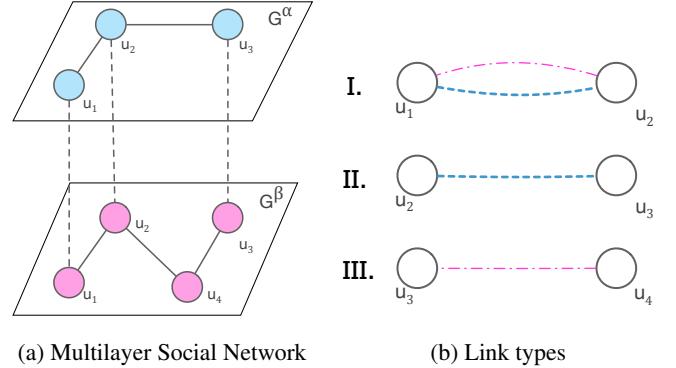
The network of human interactions is usually represented by a graph G where the nodes in set V represent people and the edges E represent interactions. While this representation has been immensely helpful for the uncovering of many social phenomena, it is focused on a single-layer abstraction of human relations. In this section, we describe a model, which represents the multiplexity of OSNs by supporting multiple friendship and interaction links.

Multilayer Online Social Network

We represent the parallel interactions between nodes across OSNs as a *multilayer network* \mathcal{M} , an ensemble of M graphs, each corresponding to an OSN. We indicate the α -th layer of the multilayer as $G^\alpha(V^\alpha, E^\alpha)$, where V^α and E^α are the sets of vertices and edges of the graph G^α . We can then denote the sequences of graphs composing the M -layer multilayer graph as $\mathcal{M} = \{G^1, \dots, G^\alpha, \dots, G^M\}$. The graphs are brought together as a multilayer system by the common members across layers as illustrated in Figure 1.

Multilayer social networks are a natural representation of media multiplexity, as each layer can depict an OSN. Figure 1a illustrates the case at hand, where there are two OSN platforms represented by G^α and G^β . Members need not be present at all layers and the multilayer network is not limited to two layers. While each platform can be explored separately as a network in its own right, this does not capture the dimensionality of online social life, which spans across multiple OSNs.

Figure 1b illustrates three link types as observed in Figure 1a for the case of a two layer network. Firstly, we define



(a) Multilayer Social Network

(b) Link types

Figure 1: Multilayer model of OSNs with I. Multiplex link;

II. Single-layer link on G^α ; and III. Single-layer link on G^β .

a *multiplex link* between two nodes i and j as a link that exists between them *at least in two layers* $\alpha, \beta \in \mathcal{M}$. Second, we say that a *single-layer link* between two nodes i and j exists if the link appears *only in one layer* in the multilayer social network. In systems with more layers, multiplexity can take on a value depending on how many layers the link is present on. In the case at hand, given layer α and layer β , we denote the set of all links present in the multilayer network as $E^{\alpha \cup \beta}$, which yields the global connectivity.

We also define the set of multiplex links as $E^{\alpha \cap \beta}$ and the set of all single-layer links on layer α only as $E^{\alpha \setminus \beta}$. These multilayer edge sets can be further extended to the M layer network by considering more layers $\{1, \dots, M\}$ as part of the intersection or union of graphs. The presence of multiplex and single-layer links in the above edge sets defines the multilayer neighbourhood of nodes in the network, as expanded upon next.

The Multilayer Neighbourhood

Following our definition of a multilayer online social network, we can redefine the ego network of a node as the *multilayer neighbourhood*. While the simple node neighbourhood is the collection of nodes one hop away from the ego, the multilayer global neighbourhood (denoted by GN) of a node i can be derived by the total number of unique neighbours across layers:

$$\Gamma_{GNi} = \{j \in V^{\mathcal{M}} : e_{i,j} \in E^{\alpha \cup \beta}\} \quad (1)$$

and their global multilayer degree as:

$$k_{GNi} = |\Gamma_{GNi}| \quad (2)$$

which provides insight into the entire connectivity of nodes across layers, and can therefore be interpreted as a global measure of the immediate degree of a node. We can similarly define the core neighbourhood (denoted by CN) of a node i across layers of the multilayer network as:

$$\Gamma_{CNi} = \{j \in V^{\mathcal{M}} : e_{i,j} \in E^{\alpha \cap \beta}\} \quad (3)$$

and their core multilayer degree as:

$$k_{CNi} = |\Gamma_{CNi}| \quad (4)$$

where we only consider neighbours which exist across all layers. This simple formulation allows for powerful extensions of existing metrics of local neighbourhood similarity. We can define the overlap (Jaccard similarity) of two users i and j 's global neighbourhoods as:

$$\text{sim}_{GNij} = \frac{|\Gamma_{GNi} \cap \Gamma_{GNj}|}{|\Gamma_{GNi} \cup \Gamma_{GNj}|} \quad (5)$$

where the number of common friends is divided by the number of total friends of i and j . The same can be done for the core degree of two users. The Jaccard coefficient, often used in information retrieval, has also been widely used in link prediction (Liben-Nowell and Kleinberg 2007).

We can further extend our definition of the multilayer neighbourhood to the Adamic/Adar coefficient for link likelihood (Adamic and Adar 2001), which considers the overlap of two neighbourhoods based on the popularity of common friends (originally through web pages) in a single-layer network as:

$$\text{aa_sim}_{GNij} = \sum_{z \in \Gamma_{GNi} \cap \Gamma_{GNj}} \frac{1}{\log(|\Gamma_{GNz}|)} \quad (6)$$

where it is applied to the global common neighbours between two nodes but can be equally applied to their core neighbourhoods. This metric has shown to be successful in the link prediction in its original single-layer form in both social networks and location-based networks (Liben-Nowell and Kleinberg 2007; Scellato, Noulas, and Mascolo 2011). In the present work, we aim to show its applicability to the multilayer space in predicting online social links across and between Twitter and Foursquare. We will next describe the specific datasets, which we apply this framework to.

Dataset

Twitter and Foursquare are two of the most popular social networks, both with respect to research efforts and user base. They have distinct broadcasting functionalities - microblogging and check-ins. While Twitter can reveal a lot about user interests, Foursquare check-ins provide a proxy for human mobility. In Foursquare users check-in to venues that they visit through their location enabled devices, and share their visit or opinion of a place with their friends. Foursquare is two years younger than Twitter and its broadcasting functionality is exclusively for mobile users (50M to date¹), while 80% of Twitter's 284M users are active on mobile². Twitter generally allows anyone to "follow" and be "followed", where followers and followed do not necessarily know one another. On the other hand, Foursquare supports undirected links, referred to as "friendship". A similar undirected relationship can be constructed from Twitter, where a link can be considered between two users if they both follow

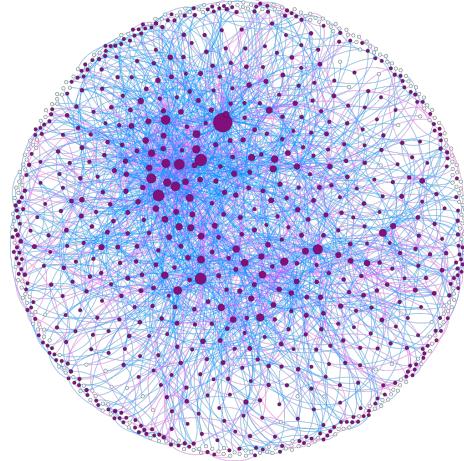


Figure 2: Social network graph for San Francisco. Blue edges are single-layer edges, while pink edges are multiplex edges. The node size is proportional to the global degree of that node.

each other reciprocally (Kwak et al. 2010). Since we are interested in ultimately in predicting friendship, we consider only reciprocal Twitter links throughout this work.

Our dataset was collected from Twitter and Foursquare in the United States between May and September 2012, where tweets and check-ins were downloaded for users who had checked-in during that time, and where those check-ins were shared on Twitter. This allows us to study the intersection of the two networks through a subset of users who have accounts and are active on both Twitter and Foursquare, and have chosen to share their check-ins to Twitter.

Property	New York	Chicago	SF	All
$ V^M $	6,401	2,883	1,705	10,989
$ E^{T \cap F} $	9,101	5,486	1,517	16,104
$ E^{T \setminus F} $	13,623	7,949	1,776	23,348
$ E^{F \setminus T} $	6,394	4,202	863	11,459
$\langle k_{GN} \rangle$	4.55	6.12	2.44	4.63
$\langle k_{CN} \rangle$	1.42	1.9	0.89	1.47
<i>tweets</i>	2,509,802	1,288,865	632,780	4,431,447
<i>checkins</i>	228,422	105,250	46,823	380,495
<i>venues</i>	24,110	11,773	6,934	42,817

Table 1: Dataset properties: number of users (nodes); number of multiplex links (edges); number of Twitter and Foursquare only edges; average global and core degrees; activity and venues per city.

We focus our analysis on the top three cities in terms of activity during the period. Table 1 shows the details for each city, in terms of activity and venues, multilayer edges and degrees for each network, where $E^{T \cap F}$ denotes the set of edges, which exist on both Twitter and Foursquare, $E^{T \setminus F}$ and $E^{F \setminus T}$ are the sets of edges on Twitter only and Foursquare only respectively.

Figure 2 additionally illustrates the case of San Francisco, where blue edges represent single-layer links on ei-

¹<https://foursquare.com/about>

²<https://about.twitter.com/company>

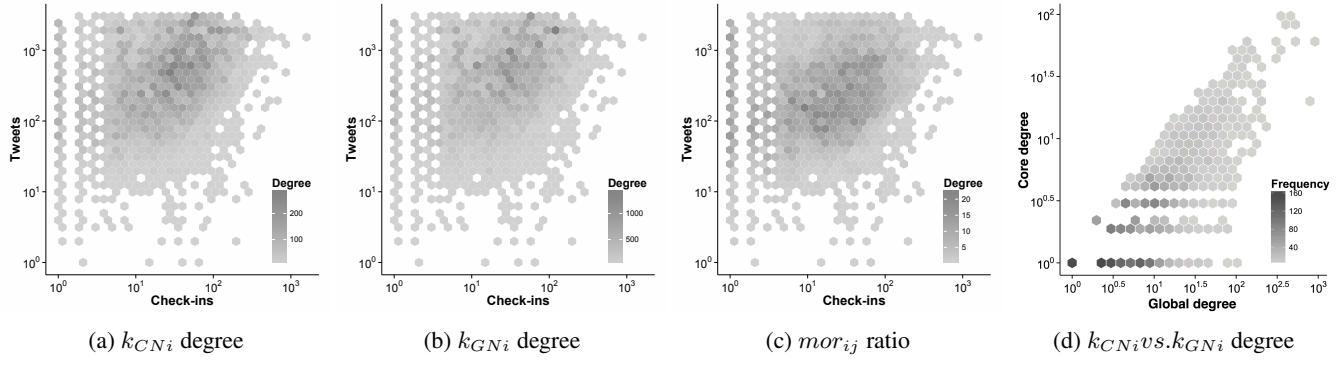


Figure 3: Multilayer degrees of users in comparison to each other and to activity volume on both networks.

ther Foursquare or Twitter, and pink edges represent multiplex links on both. We use a Fruchterman Reingold graph layout (Fruchterman and Reingold 1991) to show the core-periphery structure of the network, with larger nodes having a higher global degree k_{GN} . In the following section, we discuss the implications of these sets in detail, where we consider all three cities together, and later evaluate each one separately.

Multilayer Analysis

We begin our analysis by exploring the intersection between the Twitter and Foursquare social networks. We observe user the degree properties across the two networks at a larger scale for all three cities, while later we perform our evaluation on each city separately.

RQ1: Multilayer Degrees

We introduced two degree metrics based on the multilayer neighbourhood of a node in Equations 2 and 4, where the *global neighbourhood* is equivalent to the union of neighbours on both networks, and the *core neighbourhood* is equivalent to the intersection of neighbours across both networks. In this section we consider how the degrees relate to user activity and each other.

In both cases (Figures 3a and 3b), users with high activity on both networks, and in particular with high Twitter activity, have the highest degrees in both the core and global neighbourhoods. When we compare the two in Figure 3d, we observe that their joint distribution follows the long-tail exhibited in single-layer social networks as well. Further, we observe the multiplex overlap ratio of the core to global neighbourhood degrees in Figure 3c. This is simply the core over the global degree:

$$mor_i = \frac{k_{CNi}}{k_{GNi}} \quad (7)$$

which indicates the percent of multiplex links in i 's multilayer neighbourhood. High activity nodes across both layers at the centre of Figure 3c have the highest overlap.

In Figure 3d, we compare the two multilayer degrees. We note that the majority of users have a low degree in both,

and there is a relationship between the two. The core degree is bound by the global degree and is always a fraction of it, while the global degree may never exceed the sum of the individual layer degrees. This relationship is apparent in the figure, where *the highest degree users are those who have a large number of links which overlap (multiplex links)*. This can be due to the fact that these users are more engaged across the two platforms. We further explore the value of link multiplexity in the following section.

RQ2: Link Multiplexity

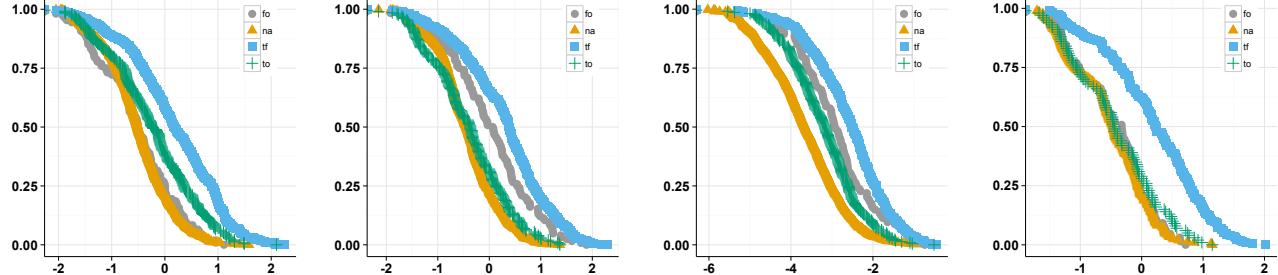
We study the three types of links as described in our multilayer model above: multiplex links on both Twitter and Foursquare, which we denote as *tf* for simplicity; single-layer links on Foursquare only (denoted as *fo*); single-layer links on Twitter only (denoted as *to*), and compare these to unconnected pairs of users (denoted as *na*). We consider reciprocal Twitter links only, where $e_{ij}, e_{ji} \in E^T$. Reciprocal relationships in Twitter have been considered as equivalent to undirected ones in other OSNs (Kwak et al. 2010).

Multiplexity and Neighbourhoods

The number of common friends has been shown to be an important indicator of a link in social networks (Liben-Nowell and Kleinberg 2007). Moreover, the neighbourhood overlap weighted on the popularity of common links between two users has been shown to be a good predictor of friendship in online networks (Adamic and Adar 2001). Figure 4 shows the Adamic/Adar metric of neighbourhood similarity across the various single and multilayer neighbourhoods described in Section 3, and the four link types.

The Adamic/Adar metric is distinctly higher for multiplex links. In agreement with previous studies of tie strength (Gilbert and Karahalios 2009), we observe that multiplex links share a greater overlap in all single and multilayer neighbourhoods. In single-layer neighbourhoods (Figure 4a and 4b) we observe that after multiplex links, those links internal to the network under consideration have a higher overlap than exogenous ones (*to* in Figure 4a and *fo* in Figure 4b), followed by unconnected pairs, which have the least overlap.

With respect to the multilayer neighbourhoods, we can observe a much more pronounced overlap across the link



(a) Twitter Adamic/Adar overlap (b) Foursq. Adamic/Adar overlap (c) Γ_{GNij} Adamic/Adar overlap (d) Γ_{CNij} Adamic/Adar overlap

Figure 4: CCDF function of the log Adamic/Adar metric for the different neighbourhoods between the four link types.

types. While the global neighbourhood overlap follows a similar distribution to the single-layer neighbourhoods but at a much lower scale, in Figure 4d we can observe more clearly that unconnected pairs share little if any neighbours, while multiplex links have a significant overlap. With respect to the global neighbourhood (Figure 4c), both Foursquare only and Twitter only links share significantly more overlap (scale is higher on x axis) than when observing the single-layer neighbourhoods in Figures 4a and 4b. This indicates that some common neighbours lie across layers, and not just within, *the global neighbourhood revealing a more complete image of connectivity, which stretches beyond the single network*.

The core neighbourhood overlap is most prominent for multiplex links (Figure 4d), which indicates that they share more friends across networks than any other type of link. While this is expected, it *confirms that the neighbourhood overlap is a good indicator of multiplexity in ties*, and is particularly strengthened in its weighted form through the Adamic/Adar metric of neighbourhood similarity.

Multiplexity and Interaction

The volume of interactions between users is often used as a measure of tie strength (Onnela et al. 2007). In this section we compare how the volume of interactions reflects on multiplex and single-layer links. We consider the following interactions on Twitter and Foursquare:

Number of mentions: This interaction feature simply measures the number of times user i has mentioned user j on Twitter during the period. Any user on Twitter can mention any other user and need not have a directed or undirected link to the user he is mentioning.

Number of common hashtags: Similarity between users on Twitter can be captured through common interests. Topics are commonly expressed on Twitter with hashtags using the # symbol. Similar individuals have been shown to have a greater likelihood of forming a tie through the principles of homophily (McPherson, Smith-Lovin, and Cook 2001).

Number of colocations: The number of times two users have checked-in to the same venue within a given time window. In order to reduce false positives, we consider a shorter time window of 1 hour only. Two users at the same place, at the same time on multiple occasions, increases the like-

lihood of them knowing each other (and having a link on social media). We weight each colocation on the popularity of a place in terms of the total user visits, to reduce the probability that colocation is by chance at a large hub venue such as an airport or train station.

Distance: Human mobility and distance play an important role in the formation of links, both online and offline, and have been shown to be highly indicative of social ties and useful for link prediction (Wang et al. 2011). We calculate the distance between the geographic coordinates of two users' most frequent check-in locations as the Haversine distance, the most common measure of great-circle spherical distance:

$$dist_{ij} = \text{haversine}(lat_i, lon_i, lat_j, lon_j) \quad (8)$$

where the coordinate pairs for i, j are of the places where those users have checked-in most frequently, equivalent to the mode in the multiset of venues where they have checked-in. We only consider users who have more than two check-ins over the whole period, and resolve ties by picking an arbitrary venue location from the top ranked venues of a user.

In Figures 5a to 5d, we observe four types of geographic and social interaction on the two social networking services, where each box-and-whiskers plot represents an interaction between multiplex links (tf), Twitter only (to), Foursquare only (fo), and unconnected pairs (na) on the x axis. On the y axis we can observe the distribution in four quartiles, representing 25% of values each. The dark line in the middle of the box represents the median of the distribution, while the dots are the outliers. The “whiskers” represent the top and bottom quartiles, while the boxes are the middle quartiles of the distribution.

In terms of Twitter mentions (Figure 5a), multiplex ties and non-connected pairs of users exhibit an overall greater number of mentions than any other group, including the Twitter only group. It is uncommon that pairs connected on Foursquare only mention each other. Mentions are quite common between users who are not connected on any network, which may be as a result of mentioning celebrities and other commercial accounts. This is not the case for hashtags, where we find that almost all of unconnected users share 10 or less hashtags with the exception of outliers. Hashtags distinguish the link type between users better than mentions.

With regards to Foursquare interaction, multiplex ties have the highest probability of multiple colocations, with Foursquare and Twitter only ties having less, and unconnected pairs more so with the exception of some outliers. In terms of distance, Twitter only and unconnected pairs are the furthest apart in terms of most frequented location, making multiplex and Foursquare links more distinguishable through this feature, as those pairs have less distance between their most frequented locations.

Although *there is certainly greater interaction between multiplex links*, followed by Twitter only and Foursquare only links, we would like to eliminate the randomness introduced by the positive results for unconnected pairs (*na*). We propose two multilayer interaction metrics combining heterogenous features from both networks in order to better distinguish between the different link types. Firstly, we define the global similarity as the Twitter similarity over Foursquare distance as:

$$sim_{GNij} = \frac{sim_{ij}^a}{dist_{ij}^b} \quad (9)$$

where *sim* can be replaced with any type of similarity, which is the mass or sum of that similarity for a pair of users, and *a, b* are exponents which can be tuned to optimise the features. Figure 5f shows how this feature captures the different levels of links (*a=2, b=1*). We additionally frame a feature which captures the complete interaction across layers of social networks:

$$int_{GNij} = \sum_{\alpha}^M |int_{ij}^{\alpha}| \quad (10)$$

where *int* can be any type of interaction of layer α , this can be further refined by giving a weight to each interaction but in our case, we consider the coefficient to be equal to 1 and use colocations from the Foursquare layer and mentions from the Twitter layer to express the global interaction of two users in the multilayer network. This feature allows us to capture the levels of different link types significantly better as shown in Figure 5e.

Although we base our analysis on only two of many possible communication channels online, we are nonetheless able to observe the greater overlap of neighbourhoods and higher intensity of interaction characteristic of multiplex links, which is in consistency with the theory of media multiplexity (Haythornthwaite 2005). We evaluate the predictive performance of the union of the features presented in the following section.

Multiplexity & Link Prediction

In this section we address the link prediction problem across layers of social networks, and aim to answer our final two research questions: *Can we predict one network using information from the other?*, and *Can we predict multiplex links in OSNs?* We evaluate the likelihood of forming a social tie as a process that depends on a union of factors, using the Foursquare, Twitter, and multilayer features we have defined

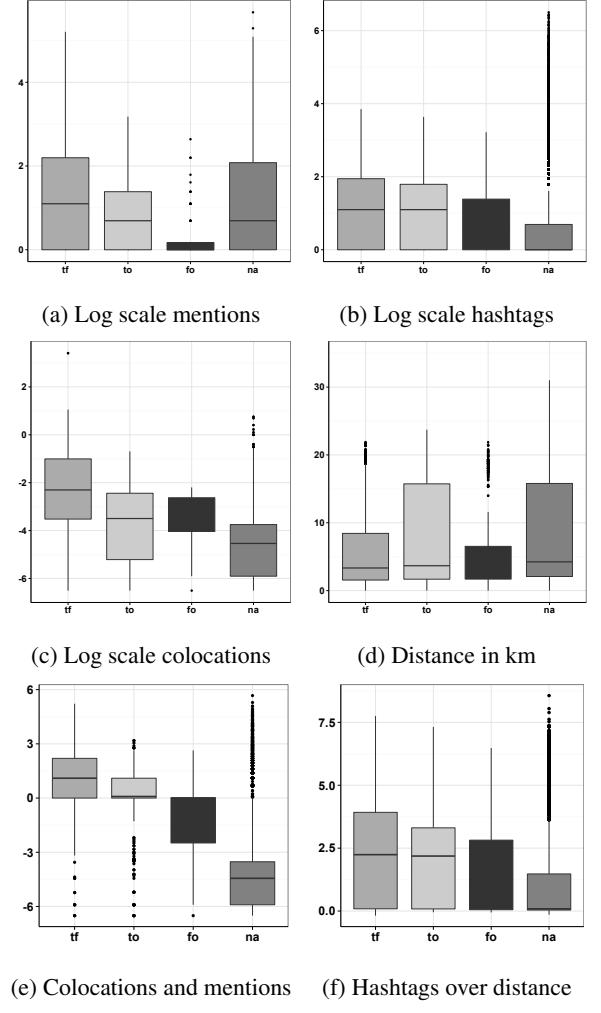


Figure 5: Interaction features for the different link types.

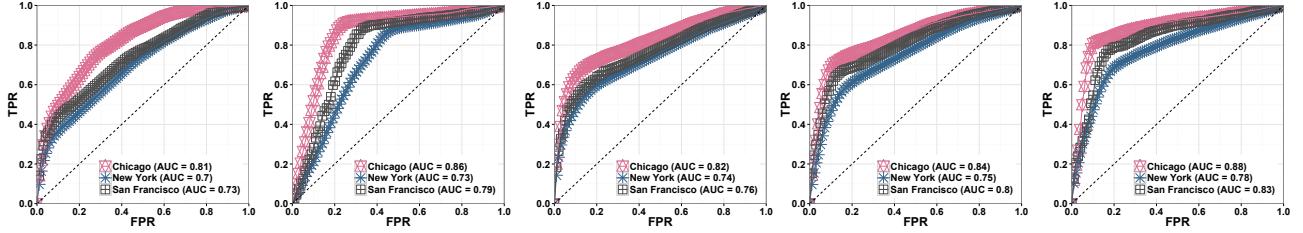
up until now in a supervised learning approach, and comparing their predictive power in terms of AUC scores for each feature set respectively.

Prediction Space

The main motivation for considering multiple social networks in a multilayer construct is that each layer carries with it additional information about the links between the same users, which can potentially enhance the predictive model. In light of the multilayer nature of OSNs, we are also interested in whether we can achieve better prediction by combining features from multiple networks.

Formally, for two users $i, j \in \mathcal{M}$, where $V^{\mathcal{M}}$ are the nodes (users) that are present in any layer of the multilayer network, we employ a set of features that output a score r_{ij}^{α} so that all possible pairs $V^{\mathcal{M}} \times V^{\mathcal{M}}$ are ranked according to their expectation of having a link e_{ij}^{α} on a specific layer α in the network. We specify and evaluate two distinct prediction tasks.

Our first goal is to rank pairs of users based on their inter-



(a) Foursq. link prediction (b) Twitter link prediction (c) Twitter feature set (d) Foursquare feature set (e) Multilayer feature set
Figure 6: ROC curves for the Random Forest classifier and Area Under the Curve (AUC) scores for each city dataset.

action on one social network in order to predict a link on the other. This entails using mobility interactions to predict social links on Twitter, and using social interactions on Twitter to predict links on Foursquare. Subsequently, we are interested in predicting the multiplex links at the cross-section of the two networks using multilayer features. This type of links have both structural and social tie implications as we have demonstrated in this work, which makes them desirable to identify.

We perform our evaluation on three datasets described at the start of this work in Section 5, where we have Twitter, Foursquare, and the derived multilayer features for the cities of San Francisco, Chicago, and New York. We adopt a supervised learning approach for the prediction tasks, and for each city, which is considered as an independent multilayer network, where we train and test on different layers. Supervised learning methodologies have been proposed as a better alternative to unsupervised models for link prediction (Lichtenwalter, Lussier, and Chawla 2010).

We compare the performance of feature sets using the Random Forest classifier (Breiman 2001) with 10-fold cross-validation testing strategy: for each test we train on 90% of the data and test on the remaining 10%. For every test case the user pairs in the test set were ranked according to the scores returned by the classifiers for the positive class label (i.e., for an existing link), and subsequently, Area Under the Curve (AUC) scores were calculated by averaging the results across all folds. We use AUC scores as a measure of performance because it considers all possible thresholds of probability in terms of true positive (TP) and false positive (FP) values rate, which are computed by comparing the predicted output against the target labels of the test data.

In terms of algorithmic implementation, we have used public versions of the algorithms available in (Pedregosa et al. 2011). The features presented earlier in this work, of which each feature set comprises are summarised in Table 2. We denote the Twitter neighbourhood as Γ^T and the Foursquare neighbourhood as Γ^F . Next, we specify each prediction task and present the results of the supervised learning evaluation in terms of the predictive power of each feature set in both tasks.

Twitter features	
<i>mentions</i>	$ mentions_{ij} $
<i>hashtags</i>	$ hashtags_{ij} $
<i>overlap</i>	$\frac{ \Gamma_i^T \cap \Gamma_j^T }{ \Gamma_i^T \cup \Gamma_j^T }$
<i>aa_sim</i>	$\sum_{z \in \Gamma_i^T \cap \Gamma_j^T} \frac{1}{\log(\Gamma_z^T)}$
Foursquare features	
<i>colocs</i>	$ colocations_{ij} $
<i>dist</i>	$haversine(lat_i, lon_i, lat_j, lon_j)$
<i>overlap</i>	$\frac{ \Gamma_i^F \cap \Gamma_j^F }{ \Gamma_i^F \cup \Gamma_j^F }$
<i>aa_sim</i>	$\sum_{z \in \Gamma_i^F \cap \Gamma_j^F} \frac{1}{\log(\Gamma_z^F)}$
Multilayer features	
<i>int_{GNij}</i>	$\sum_{\alpha}^M int_{ij}^{\alpha} $
<i>sim_{GNij}</i>	$\frac{sim_{ij}^a}{dist_{ij}^b}$
<i>overlap</i>	$\frac{ \Gamma_{CNi} \cap \Gamma_{CNj} }{ \Gamma_{CNi} \cup \Gamma_{CNj} }$
<i>aa_sim</i>	$\sum_{z \in \Gamma_{CNi} \cap \Gamma_{CNj}} \frac{1}{\log(\Gamma_{CNz})}$

Table 2: Summary of link features.

RQ3: Cross-network prediction

The Receiver Operating Characteristic (ROC) curves (defined as the True Positive versus False Positive Rate for varying decision thresholds) and the corresponding Area Under the Curve (AUC) scores are shown in Figure 6 for the three datasets. We now discuss these results with respect to each task. In the first prediction task, for a pair of users i and j we define a feature vector x_{ij}^{α} encoding the values of the users' feature scores on layer α in the multilayer network. We also specify a target label $y_{ij}^{\beta} \in \{-1, +1\}$ representing whether the user pair is connected on the β layer under prediction.

We use the supervised Random Forest classifier (45 trees, optimised with tree depth = 25) to predict links from one layer using features from the other. Figure 6a shows the ROC curves and respective AUC scores for each dataset in predicting Foursquare links from Twitter features, ranging between 0.7 for the New York dataset to 0.81 for Chicago, and 0.73 for San Francisco. We compare this to the reverse

task of predicting Twitter links using Foursquare features in Figure 6b, where we obtain AUC scores of 0.86, 0.73, and 0.79 for the three cities respectively. We observe slightly higher results for Twitter links, and we note that this may be as a result of the higher number of Twitter links in our dataset or as a result of the greater difficulty of the inverse task.

RQ4: Multiplex link prediction

In our second prediction task, we are interested in evaluating the performance of each feature set in predicting link multiplexity. Given a feature vector x_{ij} , we would like to predict a target label $y_{ij} \in \{-1, +1\}$, where a link exists on both layers (+1) or none (-1). We compare performance of the multilayer features to the Twitter and Foursquare sets.

In this task, we use all three feature sets to predict multiplex links, which generally exhibit signs of a stronger online bond through interaction and structural properties as we have seen in the first part of this work. In Figures 6c and 6d, we observe how Twitter and Foursquare features perform in predicting multiplex links using the Random Forest algorithm again, with the highest AUC scores of 0.82 and 0.84 for each set respectively. The Foursquare feature set performs better in terms of AUC scores but the multilayer feature set outperforms both (AUC = 0.88 for Chicago), due to its inclusion of features from different layers and cross-layer structural properties.

In conclusion, it is possible to predict links between heterogeneous social networks and to predict multiplex links spanning multiple networks using multilayer features as we have seen in our subset of users. We discuss the applications of these results in the following section.

Discussion & Conclusions

In this work we have demonstrated the structural and interaction properties of links across two online social networks and have also shown the value of multilayer features in predicting links on both Twitter and Foursquare, and multiplex links. We believe that the primary contribution is methodological, since it provides a novel framework for investigating multiplexity across different social networks. The techniques discussed in this work are general and can be potentially used to investigate other scenarios for which datasets containing information about social interactions across multiple networks are available. In this section, we discuss the implications, limitations and real-world applications of these results.

Implications

Recently, social media has been increasingly alluded to as an *ecosystem*. The parallel comes after the emergence of multiple OSNs, interacting as a system, while competing for the same resources - users and their attention. We have addressed this system aspect by modelling multiple social networks as a multilayer online social network in this work. We have also identified two extensions of the node neighbourhood. The global neighbourhood or degree gives insight into

a users' full connectivity across services, this is especially important when considering users with asymmetric activity and degree across networks since their centrality in the online ecosystem can be under or over-estimated. We additionally defined the core degree, which on the other hand reveals the intersection across networks, and therefore the stronger online ties - those relevant on multiple networks.

The strength of ties manifested through multiplexity is expressed through a greater intensity of interactions and greater similarity across attributes both the offline (Haythornthwaite 2005; Hristova, Musolesi, and Mascolo 2014), and in the online context as we have seen in this work. We have introduced a number of features, which take into consideration the multilayer neighbourhood of users in OSNs. The Adamic/Adar coefficient of neighbourhood similarity in its core neighbourhood version proved to be a strong indicator of multiplex ties. Additionally, we introduced combined features, such as the global interaction and similarity over distance, which reflect more distinctively the type of link, which exists between two users, than its single-layer counterparts. These features can be applied across multiple networks and can be flexible in their construction according to the context of the OSNs under consideration.

Limitations

Media multiplexity is fascinating from the social networks perspective as it can reveal the strength and nature of a social tie given the full communication profile of people across all media they use (Haythornthwaite 2005). Unfortunately, full online and offline communication profiles of individuals were not available and our analysis is limited to two social networks. Nevertheless, we have observed some evidence of media multiplexity manifested in the greater intensity and structural overlap of multiplex links and have gained insight into how we can utilise these properties for link prediction. Certainly, considering more OSNs and further relating media multiplexity to its offline manifestation is one of our future goals, and we believe that with the further integration of social media services and availability of data this will be possible in the near future.

Our data is limited to a sub-sample of users who we know have active accounts on both networks in three US cities, Foursquare check-ins also being limited to those posted on Twitter. This excludes a number of users who may have Foursquare accounts but have not linked them on Twitter. Nevertheless, we were able to show that it is possible to predict one social network from the other in a cross-network manner and we hope to extend our prediction and analysis to a greater scale and geographical scope in the future.

Applications

Most new OSNs use contact list integration with external existing networks, such as copying friendships from Facebook through the open graph protocol.³ Copying links from pre-existing social networks to new ones results in higher social

³<https://developers.facebook.com/docs/opengraph>

interaction between copied links than between links created natively in the platform (Zhong et al. 2014). We propose that extending this copied network with a rank of relevance of contacts using multiplexity can provide even further benefits for newly launched services.

In addition to fostering multiplexity, however, new OSNs and especially interest-driven ones such as Pinterest for example, may benefit from similarity-based friend recommendations. In this work, we apply mobility features and neighbourhood similarity from Foursquare to predict links on Twitter and vice versa, highlighting the relationship between similar users across heterogeneous platforms. Similarly in (Tang, Lou, and Kleinberg 2012), the authors infer types of relationships across different domains such as mobile and co-author networks. Although using a transfer knowledge framework, and not exogenous interaction features like we do, the authors also agree that integrating social theory in the prediction framework can greatly improve results. The present work is a step towards understanding the composite nature of online social network services and hopefully towards enhancing their functionality and purpose.

Acknowledgements

This work was supported by the Project LASAGNE, Contract No. 318132 (STREP), funded by the European Commission.

References

- [Adamic and Adar 2001] Adamic, L., and Adar, E. 2001. Friends and Neighbors on the Web. *Social Networks* 25:211–230.
- [Backstrom and Leskovec 2011] Backstrom, L., and Leskovec, J. 2011. Supervised Random Walks: Predicting and Recommending Links in Social Networks. In *WSDM*. ACM.
- [Breiman 2001] Breiman, L. 2001. Random Forests. *Machine Learning* 45(1):5–32.
- [Bródka and Kazienko 2012] Bródka, P., and Kazienko, P. 2012. Multi-layered Social Networks. *CoRR* abs/1212.2425.
- [Cardillo et al. 2013] Cardillo, A.; Zanin, M.; Gomez-Gardenes, J.; Romance, M.; Garcia del Amo, A. J.; and Boccaletti, S. 2013. Modeling the Multi-layer Nature of the European Air Transport Network: Resilience and Passengers Re-scheduling Under Random Failures. *The European Physical Journal Special Topics* 215(1):23–33.
- [Crandall et al. 2010] Crandall, D. J.; Backstrom, L.; Cosley, D.; Suri, S.; Huttenlocher, D.; and Kleinberg, J. 2010. Inferring Social Ties from Geographic Coincidences. In *PNAS*, volume 107, 22436–22441.
- [Fruchterman and Reingold 1991] Fruchterman, T. M. J., and Reingold, E. M. 1991. Graph Drawing by Force-directed Placement. *Softw. Pract. Exper.* 21(11):1129–1164.
- [Gilbert and Karahalios 2009] Gilbert, E., and Karahalios, K. 2009. Predicting Tie Strength with Social Media. In *CHI*.
- [Haythornthwaite and Wellman 1998] Haythornthwaite, C., and Wellman, B. 1998. Work, Friendship, and Media Use for Information Exchange in a Networked Organization. *Journal of the American Society for Information Science* 49(12):1101–1114.
- [Haythornthwaite 2005] Haythornthwaite, C. 2005. Social Networks and Internet Connectivity Effects. *Information, Communication & Society* 8(2):125–147.
- [Hristova, Musolesi, and Mascolo 2014] Hristova, D.; Musolesi, M.; and Mascolo, C. 2014. Keep Your Friends Close and Your Facebook Friends Closer: A Multiplex Network Approach to the Analysis of Offline and Online Social Ties. *ICWSM*.
- [Kazienko et al. 2010] Kazienko, P.; Brodka, P.; Musial, K.; and Gawarecki, J. 2010. Multi-Layered Social Network Creation Based on Bibliographic Data. In *SocialCom*.
- [Kivelä et al. 2014] Kivelä, M.; Arenas, A.; Barthelemy, M.; Gleeson, J. P.; Moreno, Y.; and Porter, M. A. 2014. Multilayer networks. *Journal of Complex Networks* 2(3):203–271.
- [Kwak et al. 2010] Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is Twitter, a Social Network or a News Media? In *WWW*, 591–600.
- [Lee et al. 2014] Lee, K.; Ganti, R. K.; Srivatsa, M.; and Liu, L. 2014. When Twitter meets Foursquare: Tweet Location Prediction using Foursquare. In *UbiComp*.
- [Liben-Nowell and Kleinberg 2007] Liben-Nowell, D., and Kleinberg, J. 2007. The Link-prediction Problem for Social Networks. *Journal of the American Society for Information Science and Technology* 58(7):1019–1031.
- [Lichtenwalter, Lussier, and Chawla 2010] Lichtenwalter, R. N.; Lussier, J. T.; and Chawla, N. V. 2010. New Perspectives and Methods in Link Prediction. In *KDD*, 243–252. ACM.
- [McPherson, Smith-Lovin, and Cook 2001] McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* 27:215–444.
- [Menon and Elkan 2011] Menon, A. K., and Elkan, C. 2011. Link Prediction via Matrix Factorization. In *Machine Learning and Knowledge Discovery in Databases*. Springer. 437–452.
- [Onnela et al. 2007] Onnela, J.-P.; Saramäki, J.; Hyvönen, J.; Szabó, G.; Lazer, D.; Kaski, K.; Kertész, J.; and Barabási, A.-L. 2007. Structure and tie strengths in mobile communication networks. *PNAS* 104(18):7332.
- [Ottino et al. 2014] Ottino, R.; Casas, D. L.; Pesce, J. P.; Jr., W. M.; Wilson, C.; Mislove, A.; and Almeida, V. 2014. Of Pins and Tweets: Investigating How Users Behave Across Image- and Text-Based Social Networks. In *ICWSM*.
- [Padgett and Mclean 2006] Padgett, J. F., and Mclean, P. D. 2006. Organizational Invention and Elite Transformation: The Birth of Partnership Systems in Renaissance Florence. *American Journal of Sociology* 111:1463–1568.
- [Pedregosa et al. 2011] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* 12:2825–2830.
- [Sadilek, Kautz, and Bigham 2012] Sadilek, A.; Kautz, H.; and Bigham, J. P. 2012. Finding Your Friends and Following Them to Where You Are. In *WSDM*.
- [Scellato, Noulas, and Mascolo 2011] Scellato, S.; Noulas, A.; and Mascolo, C. 2011. Exploiting Place Features in Link Prediction on Location-Based Social Networks. In *KDD*.
- [Stopczynski et al. 2014] Stopczynski, A.; Sekara, V.; Sapiezynski, P.; Cuttone, A.; Madsen, M. M.; Larsen, J. E.; and Lehmann, S. 2014. Measuring Large-Scale Social Networks with High Resolution. *PLoS ONE* 9.
- [Szell, Lambiotte, and Thurner 2010] Szell, M.; Lambiotte, R.; and Thurner, S. 2010. Multirelational Organization of Large-scale Social Networks in an Online World. *PNAS* 107(31):13636–13641.

[Tang, Lou, and Kleinberg 2012] Tang, J.; Lou, T.; and Kleinberg, J. 2012. Inferring Social Ties Across Heterogenous Networks. In *WSDM*.

[Wang et al. 2011] Wang, D.; Pedreschi, D.; Song, C.; Giannotti, F.; and Barabasi, A.-L. 2011. Human Mobility, Social Ties, and Link Prediction. In *KDD*.

[Zhong et al. 2014] Zhong, C.; Salehi, M.; Shah, S.; Cobzarenco, M.; Sastry, N.; and Cha, M. 2014. Social Bootstrapping: How Pinterest and Last.Fm Social Communities Benefit by Borrowing Links from Facebook. In *WWW*.